# Chapter 3 Data Mining Using BLAST

In the last few years, the data available for DNA and protein sequences is so enormous that searching for information is dubbed as "biological data mining". When a molecular biologist who came across a DNA or protein sequence, by comparing the sequence to all the sequences known which is kept in the database, we will be able to identify something similar where the function has already been described to the sequence and to predict potential functions for the query sequence or to help in modeling 3-D structures of a protein.

In this chapter, BLAST program is introduced first followed by information on database sequence searching, which is using data mining tool, BLAST. Then the following section will explain the procedure of submitting a nucleotide query using one of the BLASTn program.

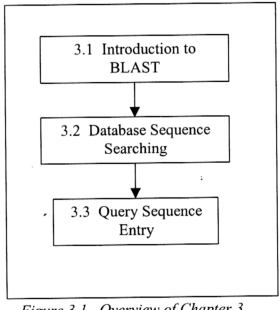


*Figure 3.1 Overview of Chapter 3*

## 3.1 Introduction to BLAST: Basic Local Alignment Search Tool

The discovery of sequence homology to a known protein or family of proteins often provides the first clues about the function of a newly sequenced gene. As the DNA and amino acid sequence databases continue to grow in size they become increasingly useful in the analysis of newly sequenced genes and proteins because of the greater chance of finding such homologies. There are a number of software tools for searching sequence databases but all use some measure of similarity between sequences to distinguish biologically significant relationships from chance similarities [19].

BLAST® (Basic Local Alignment Search Tool) is a set of similarity search programs designed to explore all of the available sequence databases regardless of whether the query is protein or DNA. The BLAST programs have been designed for speed, with a minimal sacrifice of sensitivity to distant sequence relationships. The scores assigned in a BLAST search have a well-defined statistical interpretation, making real matches easier to distinguish from random background hits. BLAST uses a heuristic algorithm, which seeks local as opposed to global alignments. A local alignment finds the optimal alignment between subregions or local regions of the specified sequence. Therefore able to detect relationships among sequences, which share, only isolated regions of similarity. [20]

BLAST is a heuristic program that attempts to optimize a specific similarity measure. It permits a tradeoff between speed and sensitivity, with the setting of threshold of parameter. The BLAST program requires time proportional to the product of the lengths of the query sequence and the database searched [21]. The program is robust and capable of analyzing both DNA and protein sequences. The BLAST server is supported through NCBI in the United States. It is the most popular, user-friendly sequence similarity search tools on the web.

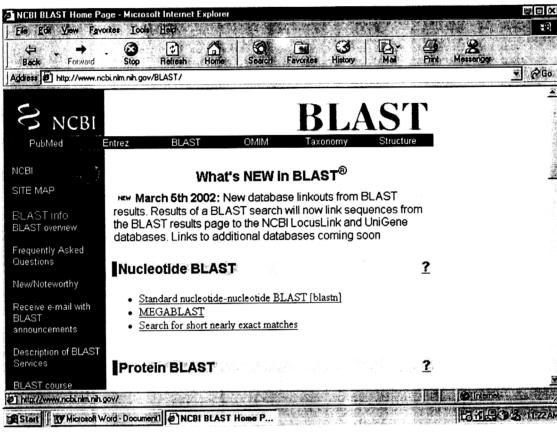


*Figure 3.2 A Basic BLAST Interface*

The core of NCBI 's BLAST services is BLAST 2.0 otherwise known as "Gapped BLAST". This service is designed to take protein and nucleic acid sequences and compare them against a selection of NCBI databases. The BLAST algorithm was written balancing speed and increased sensitivity for distant sequence relationships. BLAST is more than a tool to view sequences aligned with each other or to calculate percent homology, but a program to locate regions of sequence similarity with a view to comparing structure and function [22].

### 3.1.1 Differences of BLAST programs

BLAST is an efficient computer program for comparing DNA and protein sequences and popular software for identifying homologs of a query sequence from a database. There are several types of blast search methods and programs depending on what we have and what we want to find. The five different types of BLAST programs are:

- BLASTn: This program allows the user to search a nucleotide query sequence against a nucleotide database. A newly sequenced nucleotide query can be compared with itself or its homologs for identification and potential contamination of the query sequence.

- BLASTp: This program allows the user to search a protein query sequence against a protein database. This can be used to find all possible sequence homologs for a given protein query sequence.

- BLASTx: This program allows the user to search a translated nucleotide sequence against a protein database. The query nucleotide sequence is initially translated in all of its six possible reading frames. This program is particularly useful in finding nucleotide sequencing errors by comparing the translated nucleotide query sequence to its potential protein homologs in a protein sequence database. The information in a BLASTx output file can also help to identify unclear nucleotide in a given reported nucleotide sequence.

- tBLASTn: This program allows the user to search translated nucleotide sequences in a given nucleotide database against a protein query sequence. The nucleotide sequences in a given nucleotide database are initially translated into each of its six possible reading frames and then are compared with the protein query sequence. This program is particularly useful in finding protein sequencing errors by comparing the protein query sequence to its potential translated nucleotide homologs in a given nucleotide database. The information in a tBLASTn output

file can also help clarify unclear amino acid residues in the given query sequence. tBLASTn is similar to BLASTx in terms of its six-reading-frames translation comparison approach, but instead of a nucleotide query sequence (BLASTx) it uses a protein sequence query.

- tBLASTx: This program allows the user to search the six frame translations of a nucleotide query sequence against the six frame translations of nucleotide entries in a given nucleotide database. The tBLASTx program has similarities to both BLASTx and tBLASTn programs and can be used to complement a BLASTx search. [1]

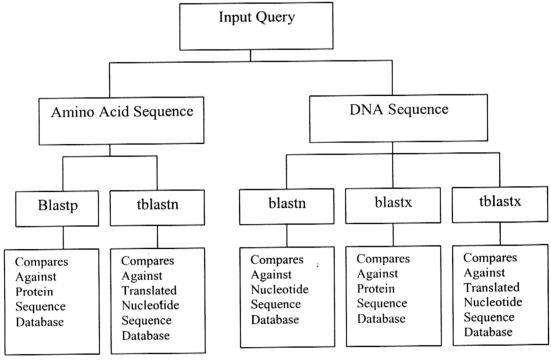


*Figure 3.3 Overview of BLAST programs*

### 3.1.2 BLASTn

In this project, the program that is used is BLASTn for searching a nucleotide sequence using fast computer algorithms to find optimal alignments of the sequences to databases. BLASTn is a standard nucleotide program that most researchers run most of the time. There are also few advance programs like MEGABLAST, PSI_BLAST and PHI-BLAST for the more complicated searches. The diagram below showed the sequence of submitting to a BLASTn program.

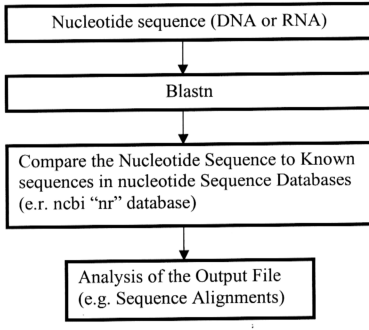


*Figure3.4 BLASTn diagram*

### 3.1.3 Databases

The National Centers used to maintain a single database that contained all the sequence data. A lot of effort went into maintaining that database (imaginatively called the non-redundant database) and removing redundant sequences from the database. A submission to any of the centers would result in the permeation of the data into all the databases. Therefore people in Europe could submit to EMBL; people in Japan to the DNA

databank of Japan; and people in the Americas to the NCBI. This is still true to some extent, and data is shared among all these systems, however because of the overwhelming amount of DNA sequence being made available, it is now not feasible to maintain separate databases. Instead there are a large, and ever growing, number of databases that we can search against. These include:

1. **nr**: All GenBank, EMBL, DDBJ, and PDB sequences (but no EST, STS, GSS, or phase 0, 1 or 2 HTGS sequences). No longer "non-redundant" but that is what the name means. It used to be the single unified database for everything.
2. **SWISS-PROT**SWISS PROT protein sequence database
3. **month**: All new or revised GenBank, EMBL (European), DDBJ (Japanese), and PDB (protein database) sequences released in the last 30 days.
4. **Drosophila genome**: Drosophila genome provided by Celera and Berkeley Drosophila Genome Project (BDGP).
5. **dbest**: Database of GenBank, EMBL, DDBJ, and PDB sequences from EST Divisions. Expressed Sequence Tags.
6. **dbsts**: Database of GenBank, EMBL, DDBJ, and PDB sequences from STS Divisions. Sequence Tagged Sites.
7. **htgs**: Unfinished High Throughput Genomic Sequences: phases 0, 1 and 2 (finished, phase 3 HTG sequences are in nr)
8. **gss**: Genome Survey Séquence, includes single-pass genomic data, exon-trapped sequences, and Alu PCR sequences.
9. **yeast**: Yeast (Saccharomyces cerevisiae) genomic nucleotide sequences
10. **E. coli**: Escherichia coli genomic nucleotide sequences
11. **pdb**: Sequences derived from the 3-dimensional structure from Brookhaven Protein Data Bank

12. **Patent**: Nucleotide sequences derived from the Patent division of GenBank

13. **vector**: Vector subset of GenBank

14. **mito**: Database of mitochondrial sequences

15. **alu**: Select Alu repeats from REPBASE (the repeat database), suitable for masking Alu repeats from query sequences.

16. **Kabat's database** of Immunologically interesting protein sequences

17. **ESTs**: ESTs from mouse, human, and other projects maintained as separate databases

18. **EPD**: Eukaryotic promoter database. [22]

## 3.3 Database Sequence Searching

The objectives of finding sequence homologs is to deduce the identity of the query sequence and to identify potential sequence homologs with known three-dimensional structures for predicting the three-dimensional structure of the target sequence and deducing functional features. A sequence database search can be started by:

1. A query sequence is needed. This is the target sequence that needs to be analyzed. The query sequence could be either a newly determined sequence whose identity is yet to be determined or one whose identity is known. A database search can help in determining the identity of the newly determined query sequence or finding possible sequence homologs for a known query sequence entry.

2. Selecting the appropriate server. The server must be reliable, regularly updated, and powerful. These characteristics are generally associated with government or government-funded bioinformatics servers such as NCBI. The National Center for Biotechnology Information (NCBI) is a collection of several public domain

databases and search tools that is readily available through the Internet and is compatible with most Web browsers.

3. Select the appropriate program or set of programs in a given server. If NCBI is the server of choice and a program is needed to conduct a simple sequence similarity search, then one of the BLAST programs might be appropriated.

4. Which BLAST program should be used for a simple sequence similarity search? If the query is a protein sequence, then BLASTp is the appropriate tool. If the query is a DNA or RNA, then the BLAST program must be utilized. These are just two of several programs available at the BLAST server. Other BLASTn programs can be utilized for finding sequence homologs for the query sequence, and also to perform more advanced tasks.
5. Select the appropriate database: There are two ways to select the appropriate database:
    - Search all relevant databases: this is a non-redundant database of all possible submitted entries. Selecting this option will enable the user to search through all the available sequence entries.
    - Search through a specific database; in this case, the user is merely interested in a specific type of database. For instance, if the user is solely interested in finding sequence homologs with known three-dimensional structures, then the PDB (Protein DataBank) database would be the most logical choice since the three dimensional structure of all its sequence entries are known.

6. Select the appropriate filter. For the convenience of its subscribers, BLAST has incorporated a set of filter options in each of its programs. The filter option excludes sequences with low complexity regions. Due to the repetitive nature of these sequences, the probability of false positive hits or random hits within the search increases and ultimately obscures the result section. It is recommended that a filter be used in a given search to reduce the number of false positives.

7. Reading, comprehending, and analyzing the output file. In order to derive a possible hypothesis from the result section of the searched query, the user needs to be familiar with the terminology used in the output file. The key subjects of the output file are its assigned score for each of the found entries, and the databases and accession numbers associated with each of those entries. The score assigned to each of the sequence found is typically an indication of its homology to the query sequence. In a BLAST output file the score is also related to the expected, or E, value assigned to each entry. The E value in a BLAST output file is the probability of the sequence being a random or chance hit. The closer it is to zero, the smaller the chance of it being a random hit from a given database. [1]

## 3.4 Query Sequence Entry

The following are the general steps a user needs to follow for a successful BLAST run:

1. The query sequence of interest must be in the correct format (e.g.,FASTA format). If the query sequence was retrieved from NCBI's Entrez, the easiest route is to copy and pasted the FASTA format of the sequence directly from Entrez into the BLAST interface.

2. The proper formatted sequence can then be pasted into the "input sequence" box on the BLAST web interface.

3. Depending on the type of sequence analyzed, the appropriate BLAST program is selected (e.g., BLASTp for protein sequences, BLASTn for DNA or RNA sequences, etc).

4. Finally, the appropriate database must be selected. The default database on BLAST is the NCBI's nr database. The nr database will search for all the available non-redundant sequences present. For example, if the user is only interested in finding sequence homologs whose structures are known, then it

would be wise to search a database that is specific to molecules with known structures. Therefore, instead of using nr, the user would select PDB as the preferred database. The sequence is now ready to be submitted to the BLAST server. The results of the search can be obtained either by e-mail or seen interactively on the BLAST web interface. The e-mail route is preferable when analyzing multiple sequence files. This allows the user to analyze the sequences of interest in a time-efficient manner, while being able to analyze the result sections later.

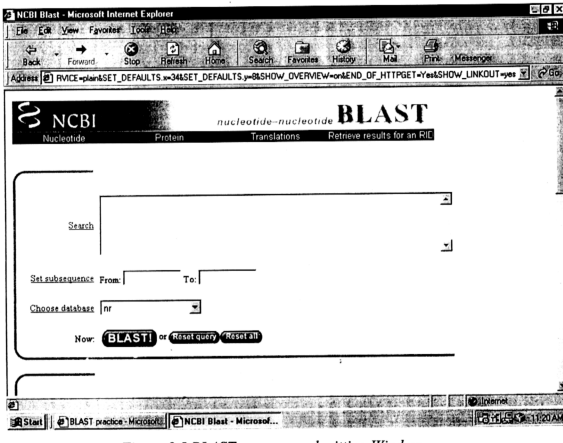


*Figure 3.5 BLAST sequence submitting Window*

## 3.5 Chapter Summary

By using the BLAST program to search for sequence similarity, useful information can be obtained to conduct further research on the function of the nucleotide search and predict the three-dimensional structure. The BLAST approach permits the construction of extremely fast programs for database searching. Given the increasing size of sequence databases, BLAST can be a valuable tool for the molecular biologist.